Outline

- Apriori improvements
- Closed frequent itemsets (CFI) & Maximal frequent itemsets (MFI)
- Beyond FIM for binary data

Thus far, FIM and ARM for binary, asymmetric data

- Binary
  - we only model the existence of an item in a transaction, e.g., t_1={A,B,C}
- Asymmetric
  - outcomes (i.e, {0,1} values) are not equally important
  - 1, i.e., the existence of an item in a transaction, is the most important

Tid	Transaction items
1	Butter, Bread, Milk, Sugar
2	Butter, Flour, Milk, Sugar
3	Butter, Eggs, Milk, Salt
4	Eggs
5	Butter, Flour, Milk, Salt, Sugar

Beyond FIM for binary data: Categorical attributes

- How to apply association analysis formulation to non-asymmetric / non-binary data ?

Gender	Level of Education	State	Computer at Home	Online Auction	Chat Online	Online Banking	Privacy Concerns
Female	Graduate	Illinois	Yes	Yes	Daily	Yes	Yes
Male	College	California	No	No	Never	No	No
Male	Graduate	Michigan	Yes	Yes	Monthly	Yes	Yes
Female	College	Virginia	No	Yes	Never	Yes	Yes
Female	Graduate	California	Yes	No	Never	No	Yes
Male	College	Minnesota	Yes	Yes	Weekly	Yes	Yes
Male	College	Alaska	Yes	Yes	Daily	Yes	No
Male	High School	Oregon	Yes	No	Never	No	No
Female	Graduate	Texas	No	No	Monthly	No	No
...	...	...	...	...	...	...	...

Example of an association rule:

{Level of Education=Graduate, Online Banking=Yes} $\rightarrow$ *{Privacy Concerns = Yes}*

Handling categorical variables

- Transform categorical attributes into (asymmetric) binary variables
- Introduce a new "item" for each distinct attribute-value pair
  - Avoid generating sets with >1 item of same attribute

Male	Female	Education = Graduate	Education = College	Education = High School	...	Privacy = Yes	Privacy = No
0	1	1	0	0	...	1	0
1	0	0	1	0	...	0	1
1	0	1	0	0	...	1	0
0	1	0	1	0	...	1	0
0	1	1	0	0	...	1	0
1	0	0	1	0	...	1	0
1	0	0	0	0	...	0	1
1	0	0	0	1	...	0	1
0	1	1	0	0	...	0	1
...	...	...	...	...	...	...	...

Beyond FIM for binary data: Continuous attributes

- How to apply association analysis formulation to non-asymmetric / non-binary data ?

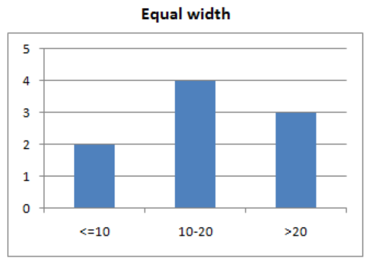
Gender	...	Age	Annual Income	No of hours spent online per week	No of email accounts	Privacy Concern
Female	...	26	90K	20	4	Yes
Male	...	51	135K	10	2	No
Male	...	29	80K	10	3	Yes
Female	...	45	120K	15	3	Yes
Female	...	31	95K	20	5	Yes
Male	...	25	55K	25	5	Yes
Male	...	37	100K	10	1	No
Male	...	41	65K	8	2	No
Female	...	26	85K	12	1	No
...	...	...	...	...	...	...

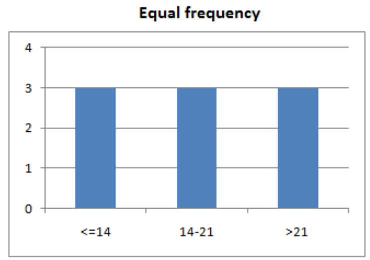
Example of an association rule:

$$\{Age \in [21,30),\ No\_of\_hours\_online \in [10,20)\} \rightarrow \{Chat\_Online = Yes\}$$

Handling continuous attributes

- Discretization
  - Equal-width binning
  - Equal-depth binning
  - …



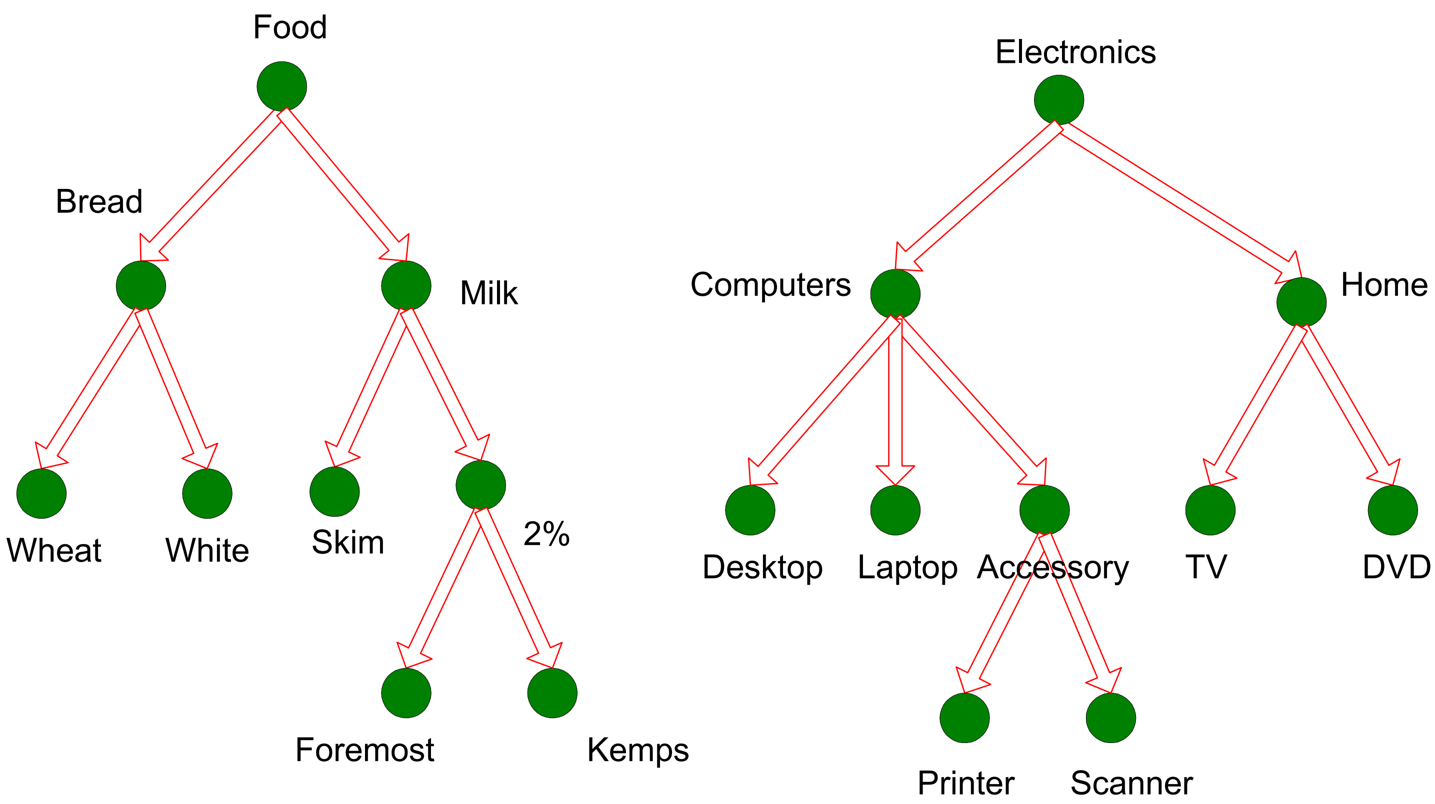


Source: http://www.saedsayad.com/images/Binning_2.png

Male	Female	$\cdots$ $\cdots$	Age < 13	Age $\in [13, 21)$	Age $\in [21, 30)$	$\cdots$ $\cdots$	Privacy = Yes	Privacy = No
0	1	$\cdots$	0	0	1	$\cdots$	1	0
1	0	$\cdots$	0	0	0	$\cdots$	0	1
1	0	$\cdots$	0	0	1	$\cdots$	1	0
0	1	$\cdots$	0	0	0	$\cdots$	1	0
0	1	$\cdots$	0	0	0	$\cdots$	1	0
1	0	$\cdots$	0	0	1	$\cdots$	1	0
1	0	$\cdots$	0	0	0	$\cdots$	0	1
1	0	$\cdots$	0	0	0	$\cdots$	0	1
0	1	$\cdots$	0	0	1	$\cdots$	0	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Multi-level frequent itemsets

- Based on concept hierarchies like



Multi-level frequent itemsets

- Why should we incorporate concept hierarchy?
  - Rules at lower levels may not have enough support to appear in any frequent itemset
  - Rules at lower levels of the hierarchy are overly specific
    - e.g., skim milk → white bread, 2% milk → wheat bread, skim milk → wheat bread, etc. are indicating an association between milk and bread, i.e., {milk}→{bread}
  - Rules at higher level of hierarchy may be too generic, e.g., {Food}→{Household items}
- Approach 1:
  - Extend current association rule formulation by augmenting each transaction with higher level items
    - Original Transaction: {skim milk, wheat bread}
    - Augmented Transaction: {skim milk, wheat bread, milk, bread, food}
- Approach 2:
  - Generate frequent patterns at highest level of concept hierarchy first
  - Then, generate frequent patterns at the next highest level of the concept hierarchy, and so on